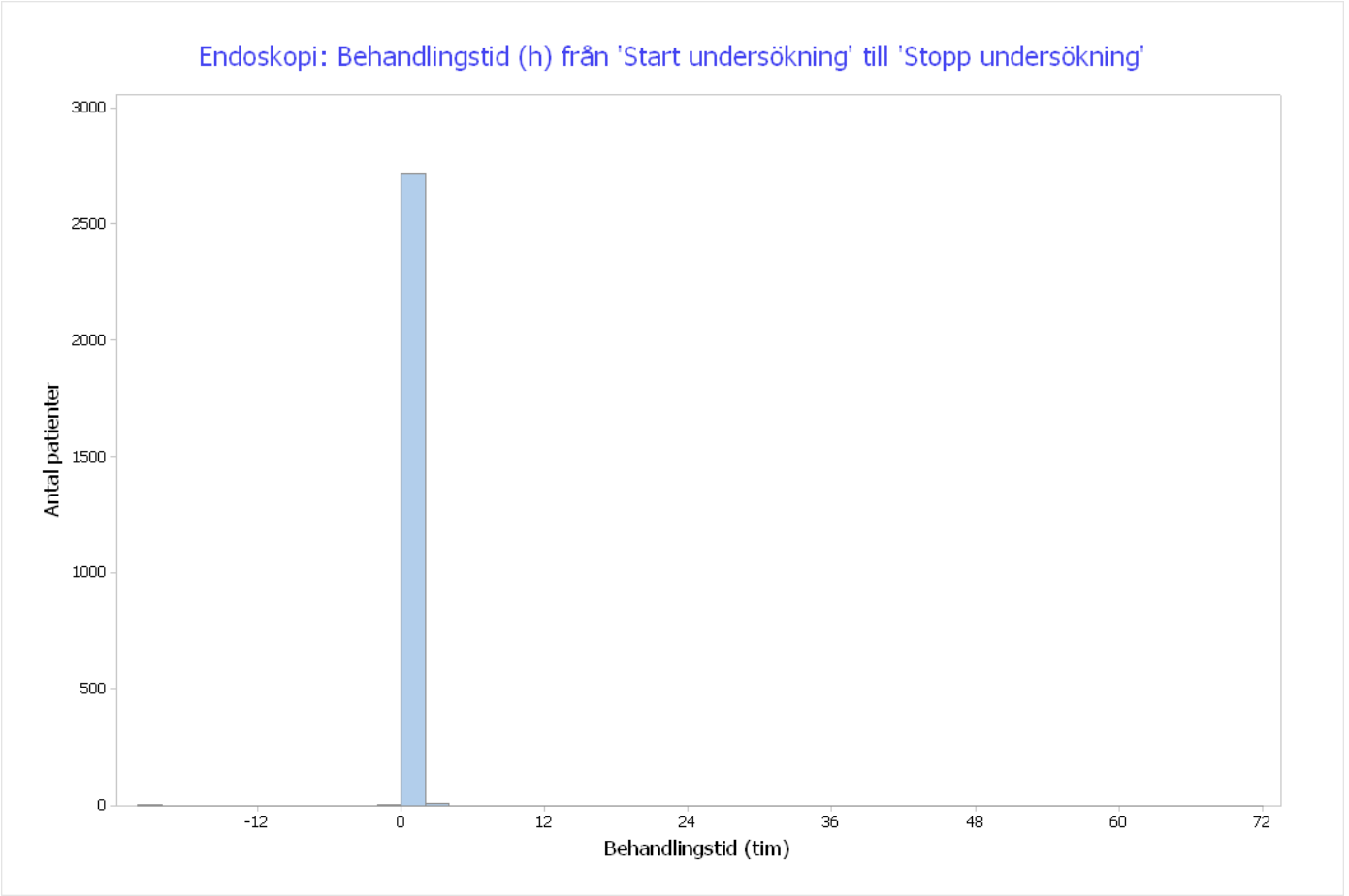
Något om datakvalitet

Lita aldrig på data!

Data från ett sjukhus – några grafer 1(3)

Tid förbehandling:

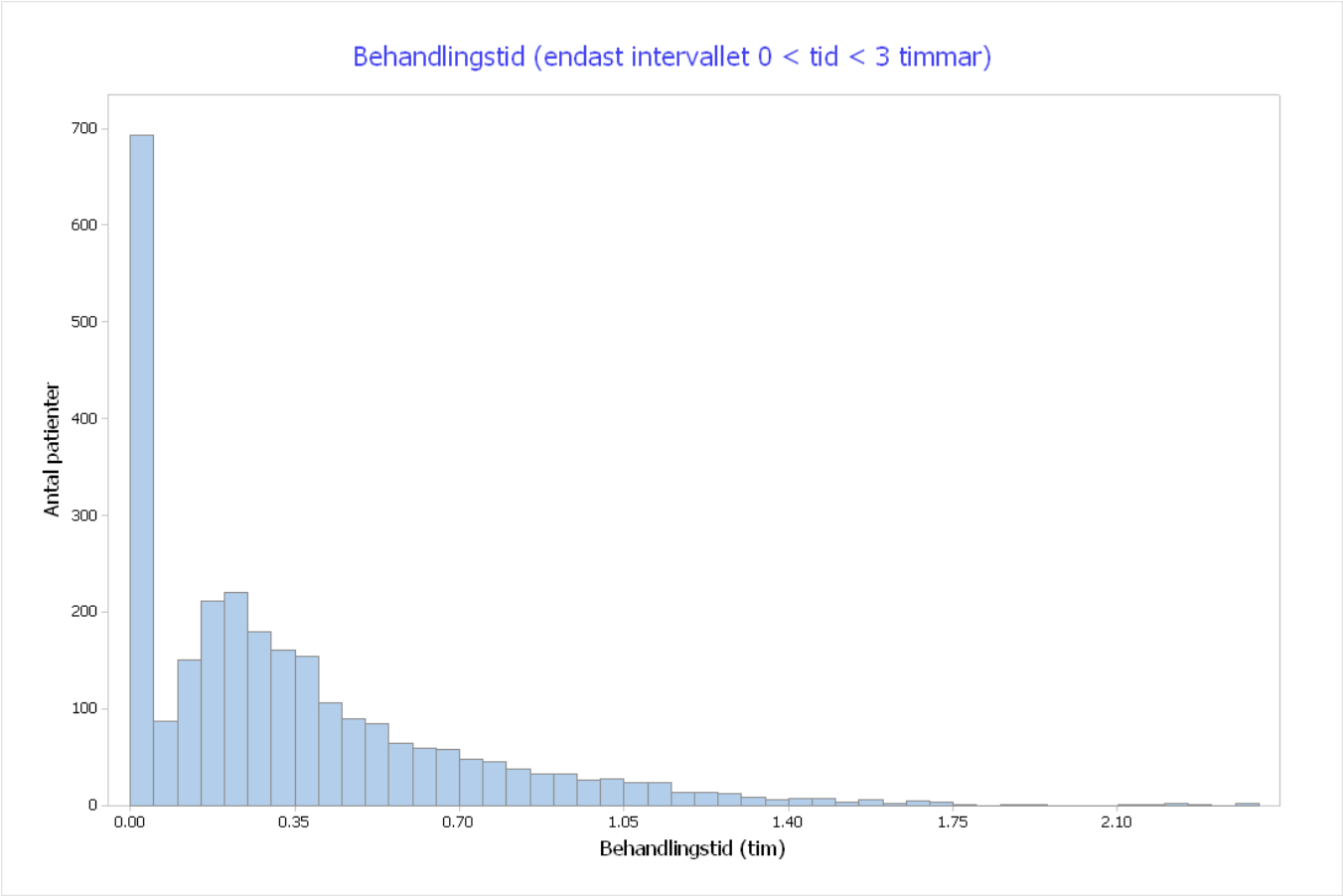
Notera negativa tider samt någon eller några extrema punkter.



Data från ett sjukhus – några grafer 2(3)

Tid förbehandling:

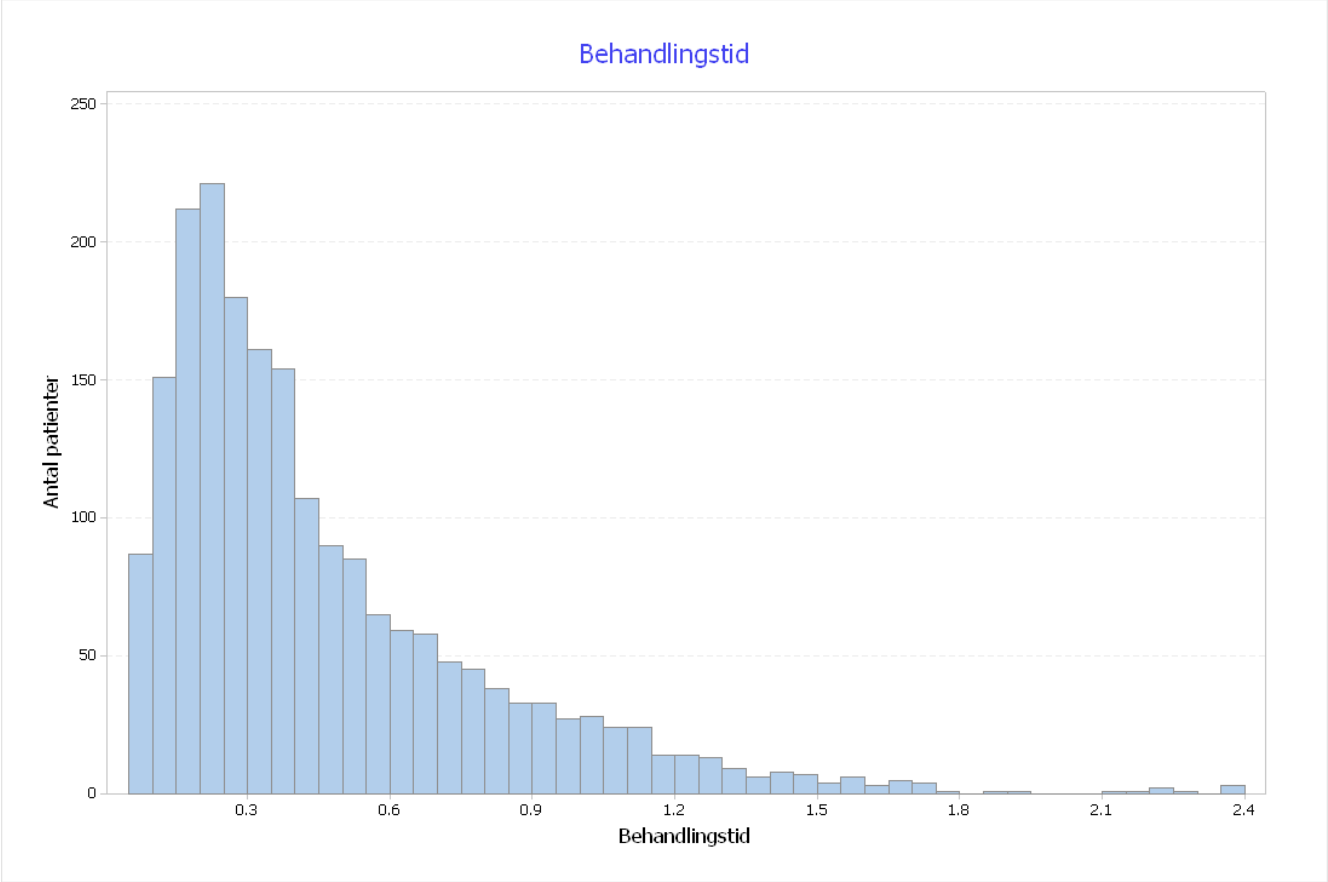
En stor mängd undersökningar går på några sekunder!



Data från ett sjukhus – några grafer 3(3)

Tid förbehandling:

Data i intervallet 3 min till 3 tim.



Några punkter om datakvalitet

- Insistera på originaldata, inte procentsatser, summor, medelvärden eller annat halvtuggat
- Kasta ’beräknade kolumner’
- Gör ’tally’ eller histogram på alla variabeltyper, koder, m.m.
- Håll reda på decimalpunkt och decimalkomma
- Håll reda på datatyp, möjliga fördelningar
- Se till att tidsangivelser (inkl datum) är givna i standardiserade format
- Undvik att transformera data
- Håll reda på korrekt uppställning i kolumner
- Kolla att kolumner har samma längd (ibland ’osynliga’ problem om data kommer t.ex. via Excel)
- Kolla visuellt på åtminstone några kolumner, rader, celler
- Gör EDA – rita många grafer, histogram, matrisdiagram, ’borstning’
- Fundera på om det behövs addition av någon slumpkomponent
- Tänk på att t.ex. ’probability plots’ är känsliga för avrundade data (Se [Statistikhörnan])

Se också
ww.ing-stat.se
knapp [Some documents] och ‘Experiences’
“Session 5 – Shaping the data,…”